Document and Network clustering with multiview point based similarity measure

Rizvana Parvin, Udhayakumar Pandian

Computer Science department, Bharath University
selaiyur, Thambaram, Chennai,TamilNadu, India

rizvanaparvin87@gmail.com

contact no: 9566022962

***Abstract* - All clustering methods have to assume some cluster relationship among the data objects that they are applied on. Similarity between a pair of objects can be defined either explicitly or implicitly. In this paper, we introduce a novel multi-viewpoint based similarity measure and two related clustering methods. The major difference between a traditional dissimilarity/similarity measure and ours is that the former uses only a single viewpoint, which is the origin, while the latter utilizes many different viewpoints, which are objects, assumed to not be in the same cluster with the two objects being measured. Using multiple viewpoints, more informative assessment of similarity could be achieved.**

***Keywords*— Document clustering, text mining, similarity measure.**

INTRODUCTION:

CLUSTERING is one of the most interesting and important topics in data mining. The aim of clustering is to find intrinsic structures in data, and organize them into meaningful subgroups for further study and analysis. Many clustering algorithms published every year. K-means is the most frequently used partitional clustering algorithm in practice. A common approach to the clustering problem is to treat it as an optimization process. An optimal partition is found by optimizing a particular function of similarity (or distance) among data. For instance, the original k-means has sum-of-squared-error objective function that uses Euclidean distance. In a very sparse and high-dimensional domain like text documents, spherical k-means, which uses cosine similarity (CS) instead of Euclidean distance as the measure, is deemed to be more suitable

PROBLEM STATEMENT:

The cosine similarity can be expressed in the following form without changing its meaning:

sim (di, dj) = cos (di-0, dj-0) = (di-0) t (dj-0)

Where 0 is vector 0 that represents the origin point. According to this formula, the measure takes 0 as one and only reference point. The similarity between two documents di and dj is determined w.r.t. the angle between the two points when looking from the origin. To construct a new concept of similarity, it is possible to use more than just one point of reference. We may have a more accurate assessment of how close or distant pair of points is, if we look at them from many different viewpoints. From a third point dh, the directions and distances to di and dj are indicated, respectively, by the difference vectors (di-dh) and (dj-dh). By standing at various reference points dh to view di, dj and working on their difference vectors, we define similarity between the two documents.

Multi-View Point Based Similarity:

Our approach in finding similarity between documents or objects while performing clustering is multi-view based similarity. It makes use of more than one point of reference as opposed to existing algorithms used for clustering text documents. As per our approach the similarity between two documents is calculated as:

Sim (di, dj) = 1/n-nr Σ Sim (di-dh, dj-dh)
dt, dj Sr dh S\Sr

Here is the description of this approach. Consider two point di and dj in cluster Sr. The similarity between those two points is viewed from a point dh which is outside the cluster. Such similarity is equal to the product of cosine angle between those points with respect to Euclidean distance between the points. An assumption on which this definition is based on is “dh is not the same cluster as di and dj. When distances are smaller the chances are higher that the dh is in the same cluster. Though various viewpoints are useful in increasing the accuracy of similarity measure there is a possibility of having that give negative result. However the possibility

of such drawback can be ignored provided plenty of documents to be clustered.

EXPERIMENTAL RESULTS AND ANALYSIS :

demonstrate how well MVSCs can perform, we compare them with five other clustering methods on the 20 data sets. In summary, the seven clustering algorithms are

. MVSC-IR: MVSC using criterion function IR
. MVSC-IV: MVSC using criterion function IV
. K-means: standard k-means with Euclidean distance
. Spkmeans: spherical k-means with CS
. graphCS: CLUTO’s graph method with CS
. graphEJ: CLUTO’s graph with extended Jaccard
. MMC: Spectral Min-Max Cut algorithm

Our MVSC-IR and MVSC-IV programs are implemented in Java. The regulating factor _ in IR is always set at 0.3 during the experiments. We observed that this is one of the most appropriate values. A study on MVSC-IR’s performance relative to different _ values is presented in a later section.

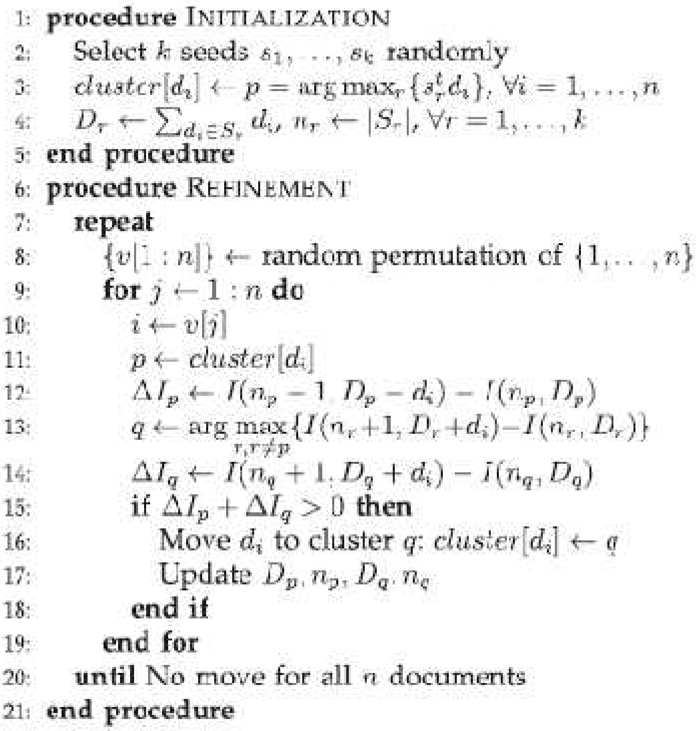
Dataset:

We are taking related datasets of construction of two bridges
Attributes specified are given below
LOCATION, ERECTED, PURPOSE, LENGTH, LANES, CLEAR, MATERIAL, SPAN, REL, TYPE
There are 13 attributes which have 7 specifications, 5 design descriptor and 1 identifier

Incremental clustering:

At Initialization, k arbitrary documents are selected to be the seeds from which initial partitions are formed. Refinement is a procedure that consists of a number of iterations. During each iteration, the n documents are visited one by one in a totally random order. Each document is checked if its move to another cluster results in improvement of the Objective function. If yes, the document is moved to the cluster that leads to the highest improvement. If no clusters are better than the current cluster, the document is not moved. The clustering process terminates when iteration completes without any documents being moved to new clusters. Unlike the traditional k-means, this algorithm is a stepwise optimal procedure. While k-means only updates after all n documents have been reassigned, the incremental clustering algorithm updates immediately whenever each document is moved to new cluster. Since every move when happens increases the objective function value, convergence
to a local optimum is guaranteed.

Algorithm:



```
procedure INITIALIZATION
    Select $k$ seeds $s_1, \ldots, s_k$ randomly
    $cluster[d_i] \leftarrow p = \arg\max_r \{s_r^t d_i\}, \forall i = 1, \ldots, n$
    $D_r \leftarrow \sum_{d_i \in S_r} d_i,\ n_r \leftarrow |S_r|, \forall r = 1, \ldots, k$
end procedure
procedure REFINEMENT
    repeat
        $\{v[1:n]\} \leftarrow$ random permutation of $\{1, \ldots, n\}$
        for $j \leftarrow 1 : n$ do
            $i \leftarrow v[j]$
            $p \leftarrow cluster[d_i]$
            $\Delta I_p \leftarrow I(n_p - 1, D_p - d_i) - I(n_p, D_p)$
            $q \leftarrow \arg\max_{r, r \neq p} \{I(n_r + 1, D_r + d_i) - I(n_r, D_r)\}$
            $\Delta I_q \leftarrow I(n_q + 1, D_q + d_i) - I(n_q, D_q)$
            if $\Delta I_p + \Delta I_q > 0$ then
                Move $d_i$ to cluster $q$: $cluster[d_i] \leftarrow q$
                Update $D_p, n_p, D_q, n_q$
            end if
        end for
    until No move for all $n$ documents
end procedure
```

CONCLUSIONS:

In this paper, we propose a Multiviewpoint-based Similarity measuring method, named MVS. Theoretical analysis and empirical examples show that MVS is potentially more suitable for text documents than the popular cosine similarity. Based on MVS, two criterion functions, IR and IV and their respective clustering algorithms, MVSC-IR and MVSC-IV, have been introduced. Compared with other state-of-the-art clustering methods that use different types of similarity measure, on a large number of document data sets and under different evaluation metrics, the proposed algorithms show that they could provide significantly improved clustering performance

The key contribution of this paper is the fundamental concept of similarity measure from multiple viewpoints. Future methods could make use of the same principle, but define alternative forms for the relative similarity or do not use average but have other methods to combine the relative similarities according to the different viewpoints. Besides, this paper focuses on partitional clustering of documents. In the future, it would also be possible to apply the proposed criterion functions for hierarchical clustering algorithms. Finally, we have shown the application of MVS and its clustering algorithms for text data. It would be interesting to explore how they work on other types of sparse and high-dimensional data.

REFERENCES:

[1] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.-H. Zhou, M. Steinbach, D.J. Hand, and D. Steinberg, "Top 10 Algorithms in Data Mining," Knowledge Information Systems, vol. 14, no. 1, pp. 1-37, 2007.

[2] I. Guyon, U.V. Luxburg, and R.C. Williamson, "Clustering: Science or Art?," Proc. NIPS Workshop Clustering Theory, 2009.

[3] I. Dhillon and D. Modha, "Concept Decompositions for Large Sparse Text Data Using Clustering," Machine Learning, vol. 42, nos. 1/2, pp. 143-175, Jan. 2001.

[4] S. Zhong, "Efficient Online Spherical K-means Clustering," Proc.

IEEE Int'l Joint Conf. Neural Networks (IJCNN), pp. 3180-3185, 2005.